Cancer Genetic Markers of Susceptibility (CGEMS) Prostate Cancer Genome-Wide Association Scan Phase 1A

The CGEMS data portal provides public access to summary results for approximately 300,000 SNPs genotyped in Phase 1A of the CGEMS prostate cancer scan of in more than 1,100 prostate cancer patients and an equivalent number of controls. The CGEMS project aims to make Phase 1B, which consists of genotypes and results for an additional 240,000 SNPs, available during the first quarter of 2007. The combination of the two phases will result in over 500,000 SNP genotypes per subject, which should provide approximately 90% coverage of common SNPs based on HapMap Phase 2 with minor allele frequency (MAF) greater than 0.05 in the European population and a linkage disequilibrium coefficient threshold of r2>0.8.

The summary data can be viewed via the CGEMS data portal and downloaded in bulk. Future data releases at this portal will include results on additional genotype data with limited phenotype information. Access to a subset of the PLCO individual raw phenotype and genotype data will be possible for research scientific purposes only after registration by the individual investigator and the supporting institution. The accessible data will include genotypes from the WGAS and a set of covariates, namely, age (in categories of 5 years, 55-60, 60-65 and 65-70), family history of cancer (yes/no), and disease phenotype (control, non-aggressive prostate cancer, aggressive prostate cancer). Access to additional covariate data will be possible through established data sharing policies of PLCO (http://www.parplco.org/; contact Danielle Carrick, PLCO EEMS Study Coordinator, Westat, Inc., Rockville, MD, tel: 240-314-5896).

Study Population

The Prostate, Lung, Colon and Ovarian (PLCO) Cancer Screening Trial is a large, study of approximately 155,000 men and women randomized to either a screening or control arm, designed to test the efficacy of cancer screening for the early detection of prostate, lung, colorectal and ovarian cancer. The protocol of the trial included collection of blood samples and questionnaire data and was designed to investigate early markers for cancer as well as the etiology of common cancers (Gohagan, 2000; Hayes, 2005). PLCO enrollment began in 1994 and ended in 2001. Recruitment included men and women, aged 55 to 74 with no history of prostate, lung, colon and ovarian cancer, although prior diagnoses of other cancers were acceptable.

The CGEMS cohort consisted of men enrolled in the screening arm of the PLCO Trial who:

1. were White and non-Hispanics;
2. had no prior history of prostate of cancer before randomization;
3. had at least one (PLCO) prostate cancer screen (PSA testing) before October 1, 2003;
4. had completed a Baseline Questionnaire about risk factors for cancer;
5. had signed informed consent;

6. had provided a blood sample of
    a. at least 11 μg DNA, or
    b. at least 1 vial of buffy coat, or
    c. at least 7 vials of whole blood was available;
7. and, for controls, had at least one Annual Study Update (ASU), a form subjects were asked to complete annually after baseline assessment to notify the study of any cancers diagnosed in the past year. For controls only, they were required to have completed at least one ASU.

Based on these criteria, 28,521 men were included in the CGEMS sub-cohort.

CGEMS distinguishes between non-aggressive and aggressive cases of prostate cancer. The two subtypes are defined as follows:

1. ***Non-aggressive***: cases with a Gleason Score < 7 and Stage < III.
2. ***Aggressive***: cases with a Gleason Sore ≥ 7 or Stage ≥ III.

All men diagnosed with non-aggressive prostate cancer between enrollment and October 1, 2001 were included for study in CGEMS. Because of our interest in the clinically more significant, aggressive form of prostate cancer, we increased the fraction of aggressive cases in the CGEMS case series by selecting all aggressive cases diagnosed with prostate cancer between enrollment and October 1, 2003. Study enrollment began on October 1, 1993. Consequently, study years in the PLCO Trial are counted according to the Federal fiscal year, Oct 1 to the following September 30.

A total of 737 aggressive cancers and 624 non-aggressive cancers met the eligibility criteria for a total of 1,361 possible subjects with prostate cancer and were considered for the CGEMS project. Of the eligible cases, all aggressive cases (n=737) were chosen to be cases in the CGEMS prostate cancer study whereas a selection of 493 (70.4%) of the possible 624 men with non-aggressive tumors patients were included in the available pool of non-aggressive cases. The selected non-aggressive cases were diagnosed earliest during follow-up time after the first screening.

To select controls by incidence-density sampling, we first created non-overlapping sets of cases with:

1. The same calendar years of entry into the cohort (October 1-September 30).
2. The same five-year age at entry groups (55-59, 60-64, 65-69, 70-74).
3. Had the same number of years under follow-up between enrollment and diagnosis of prostate cancer.

For each case set, we identified from among the 28,251 eligible men in the CGEMS Cohort those who met the following criteria:

1. Entry into the cohort in same year as the case set.
2. Same five-year age at entry (55-59, 60-64, 65-69, 70-74).
3. No prostate cancer diagnosis through at least as many years in follow-up in PLCO Trial.

For each set of cases, we initially selected a random sample of disease-free controls from the corresponding risk set with size equal to the number of cases in the set. For cases diagnosed before 2002, some samples had already been extracted for another prostate cancer study within the PLCO cohort, but all eligible subjects had equal inclusion probability for this study. We replaced some of the previously selected controls with newly enrolled cohort members according to a random selection rule that ensured that the chance of inclusion as a control in the CGEMS study for a given case would not be affected by whether the control was already selected for the previous study.

In incidence density sampling, a male subject is selected as a control for a given case set independently of eligibility and selection as a control for other case sets and independently of future diagnosis as a case. A total of 1,230 control comparison units was formed and resulted in the selection of the controls summarized in Table 1 below. 1,204 different men were identified as controls at least once (1,179 subjects sampled once, 24 subjects sampled twice, and one subject sampled three times); thus, 1230 controls were available. Forty-six control subject selections, derived from 44 eligible subjects, subsequently developed prostate cancer; 32 of these cases were included as cases and 12 were not (i.e., non-selected patients with non-aggressive prostate cancer).

Table 1. Potential CGEMS participants from PLCO cohort

Case status	Number of times selected as a control				Total
	0	1	2	3	
Not eligible as a case	26,000	1,136	23	1	27,160
Eligible, but not selected as a case	119	11	1	0	131
Eligible, and selected as a case	1198	32	0	0	1,230
All subjects	27,317	1,179	24	1	28,521

Sample handling:

DNA samples were received from the PLCO bio-repository and visually inspected for adequate fluid in individual tubes. Three measurements of quantification were performed according to the standard procedures at the Core Genotyping Facility of National Cancer Institute. These include pico-green analysis, optical density spectrophotometry and real time PCR (http://cgf.nci.nih.gov/dnaquant.cfm). Samples were also analyzed with 15 short tandem repeats and the Amelogenin marker in the Identifiler™ Assay (ABI, Foster City, CA). All samples advanced to genotype analysis completed no less than 13 of the 15 micro-satellite markers.

Table 2. Final set of PLCO samples genotyped in CGEMS

Status at initiation of CGEMS project	Number of times selected as a control				Total
	0	1	2	3	
Prostate cancer-negative during follow-up	0	1,087	22	1	1,110
Diagnosed with non-aggressive cancer	466	26	1	0	493
Diagnosed with aggressive cancer	679	16	0	0	695
All subjects	1,145	1,129	23	1	2,298

After final review and sample handling, 1188 of 1361 (87.2%) of eligible cases were genotyped in CGEMS. If the incidence density sampling strategy is ignored, there are 1,188 cases and 1,110 non-cases included in our analysis.

For quality control analysis, 49 DNAs from PLCO were genotyped in duplicate. We also genotyped 100 DNAs from CEPH families of which 73 were genotyped in duplicate.

Selection of SNPs:

Genotyping of the CGEMS Prostate Cancer Phase 1A samples was performed under contract by Illumina Corporation using the Sentrix® HumanHap300 genotyping assay. The HumanHap300 BeadChip is designed to enable the genotyping of a fixed panel of 317,000 tagSNPs identified following the method initially described by Carlson et al. (*Am J Hum Genet 2004*). This selection was performed using the data from Phase I of the International HapMap project (http://www.hapmap.org/) and a threshold for the linkage disequilibrium statistic r^2 of 0.7 for non-genic regions and 0.8 for genic regions. For the European population, this panel is expected to cover close to 90% of the common polymorphisms studied in HapMap phase 1 at an r^2 threshold of 0.8. The coverage of the HapMap phase 2 set of SNP is over 75% as evaluated by the TagZilla (http://tagzilla.nci.nih.gov/) program.

Quality control

Initial Assessment of Call Rates

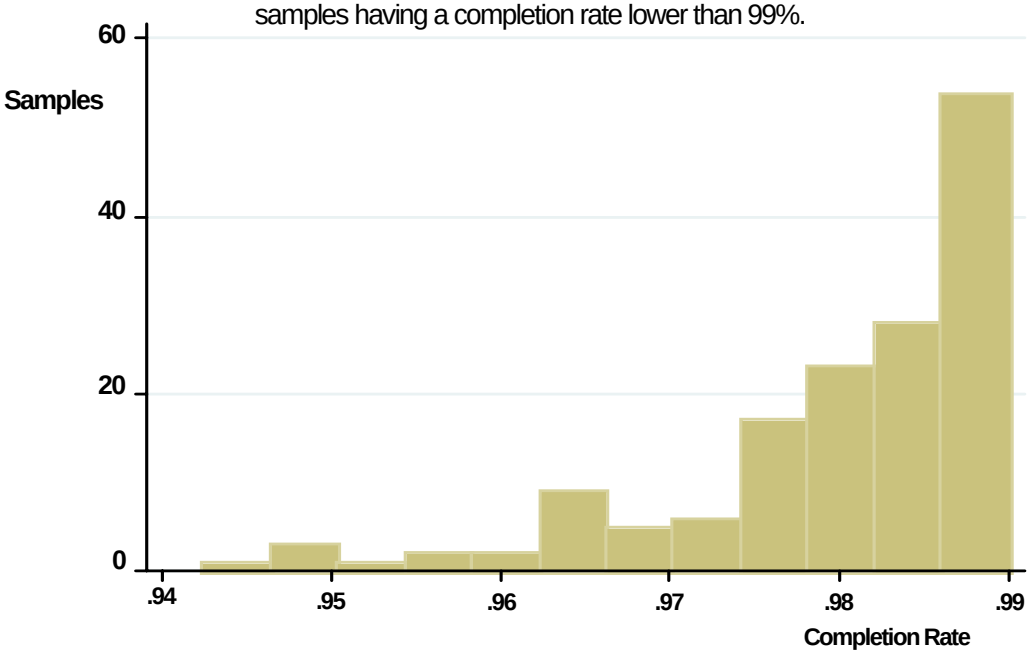
In Phase 1A, a total of 317,503 SNP genotype assays were attempted on 2,540 DNA samples. If the call rate for any sample was below 90%, the sample was excluded from further analysis; 18 CEPH samples and 7 PLCO samples were excluded by this criterion. 175 CEPH samples (including 73 duplicate DNAs including one set of a quadruplet) and 2,340 PLCO samples (including 49 duplicate DNAs) passed.

A total of 10,211 SNPs (~3.2% overall) failed to provide accurate genotype results due to either low call rates (<95%) or were mono-allelic across the study. Subsequent quality control analysis was performed on the remaining 307,292 SNPs

Completion rate for samples

Overall, 157 of 175 (90%) CEPH samples and 2,189 of 2,340 (94%) PLCO samples generated more than 304,000 high performance genotype calls (>99% completion rate). Figure 1 provides the distribution of the completion rate for the 151 remaining samples. The lowest call rate was 94.24%. Overall, the average completion rate for the PLCO samples for 307,292 SNPs is 99.7%. This rate was not significantly different in the controls compared to the combined aggressive and non-aggressive case groups (status at initiation of CGEMS project, t test p=0.30, Kruskal-Wallis p=0.46) and remained non-significant when the two case phenotypes were separated (Kruskal-Wallis p=0.76). No additional subjects were excluded from the association analysis based on completion rate.

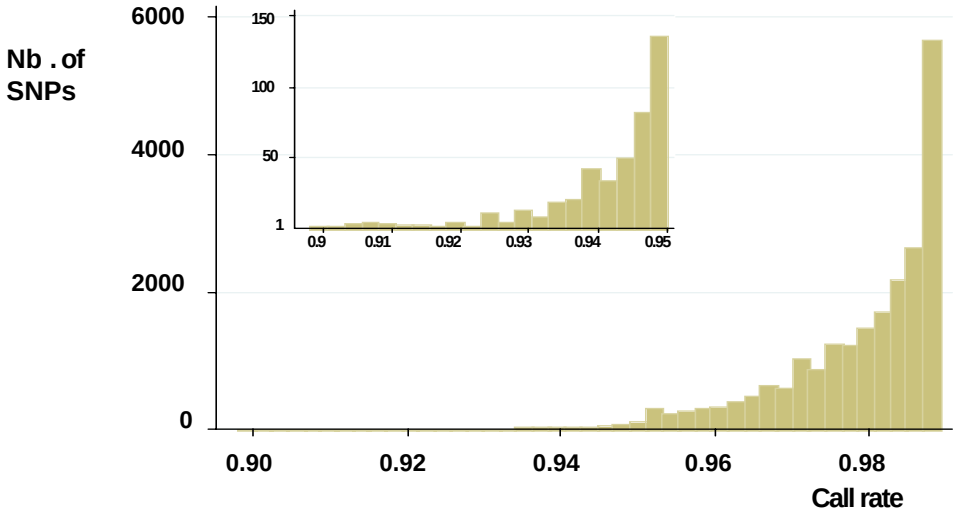
Figure1 : Distribution of the completion rates for the 151 PLCO samples having a completion rate lower than 99%.



Call rate for SNPs

A total of 285,228 (92.8%) SNPs yielded a call rate higher than 99%. Figure 2 provides the distribution of the call rate for the 22,064 remaining SNPs. The lowest call rate retained for our analysis was 89.8%.

Figure 2 Distribution of the call rates of the 22 064 SNPs having a call rate lower than 99%.



Insert distribution of call rates for the 455 SNPs having a call rate lower than 95%

Concordance rate

The genotype concordance rate for SNP assays was evaluated based on three comparisons:

1. Genotype generated on CEPH DNA.

Of the CEPH DNAs, 73 had been provided twice as separate aliquots from the same DNA preparation, then genotyped and evaluated for genotype concordance. An average of 305,000 genotypes comparison was performed for each DNA pair. Between 0 and 1,342 discordances (average 59) were observed within each pair comparison, yielding a discordance rate of $2.0 \pm 6.3 \times 10^{-4}$. Four pair of duplicates with discordance rate higher than 10^{-3} were considered outliers.

2. Genotypes generated on duplicate DNA from PLCO.

From PLCO, 49 pairs of DNA were provided twice, analyzed as separate aliquots from the same DNA preparation but performed comparably, thus, providing reliable data. Analysis of the discrepancies within these pairs revealed similar results to the CEPH DNA duplicates. An average discordance rate $2.1 \pm 5.1 \times 10^{-4}$ was observed. In both sets of DNAs the discordance rate was always lower than 4×10^{-5} when the average call rate in DNA pairs was higher than 99.6%.

3. Comparison of CGEMS and HapMap Genotypes

28 samples genotyped in CGEMS were also genotyped by the International HapMap Consortium. Comparison of the CGEMS and HAPMAP genotypes for these individuals revealed a mean discordance rate of 1.7×10^{-3}, which is an order of magnitude higher than that observed for PLCO duplicates in the CGEMS project. Notably, the International HapMap Consortium received its DNA from the Corriell Institute (Camden, New Jersey) and CGEMS received DNA from the Centre d'Etude du Polymorphisme Humain (CEPH, Paris, France). Although the DNAs were obtained from cell lines derived from the same individual, they were extracted from cell lines propagated at different times and location and processed independently. The distribution of discordances was as follows: 94% between homozygote and heterozygote calls and only 6% showed homozygote/homozygote discordances.

2,137 SNPs were removed from the analysis because of discordance rates greater than 5% by any of these comparisons.

Hardy –Weinberg Proportions in control DNA

Genotype data for all autosomal SNPs were tested for deviation from Hardy-Weinberg proportions using (Wigginton et al. *Am J Hum Genet* 76:887-893, 2005). The analysis was conducted in the PLCO control group. Significant p values ($p < 0.05$) were observed for 17,410 SNPs (5.4%). For the Phase 1A association analysis, 3,651 SNPs were removed with a p value for HW proportion lower than 0.01.

Final sample and SNP selection for association analysis
Based on the quality control analyses outlined above, 5,621 SNPs were removed from the analysis for main effect because of evidence of potentially unreliable genotype data.
Thus, a total of 301,671 SNPs were investigated for single SNP main effect association testing. Three DNA samples revealed a large number of heterozygous loci on the X chromosome, suggesting sample mix-up. In addition, three pairs of DNAs were found to be unexpected duplicates. These 9 DNAs were removed from subsequent analysis. Thus, association analysis was performed on the final set of 2,282 DNAs described in table 3

Table 3. Final set of PLCO samples analyzed for association

Status at initiation of CGEMS project	Number of times selected as a control				Total
	0	1	2	3	
Never developed prostate cancer	0	1,082	22	1	1,105
Diagnosed with non-aggressive cancer	461	26	1	0	488
Diagnosed with aggressive cancer	673	16	0	0	689
All subjects	1,134	1,124	23	1	2,282

Association Analysis

The primary analysis of the CGEMS prostate WGAS study explores the association between single SNPs and prostate cancer susceptibility. The current report is based on the completion of Phase 1A, which includes approximately 2/3 of the total Phase 1 SNPs. Phase 1B will complete the data set and the combined data will provide over 500,000 SNPs per subject.

The analytic approach assumes no structure to the risk across the 3 possible genotypes at each locus. This approach maintains power to detect recessive or over-dominant alleles at the cost of a small decrease in power relative to an Armitage trend test for the detection of alleles with multiplicative effect.

Furthermore, we distinguish between non-aggressive and aggressive prostate cancer in the analysis. Thus, our analysis is not predicated on the hypothesis that susceptibility loci have the same odds ratios for risk of developing aggressive and non-aggressive prostate cancer.

An ‘agnostic’ approach was selected to survey common genetic variation across the genome, which did not take into consideration gene function or prior implication in prostate or other cancers. Our agnostic approach maximizes genome coverage for a given number of SNPs, thereby providing a robust opportunity to pursue many different working hypotheses in the CGEMS data set and many SNPs in regions of interest now and in the future.

Case phenotypes
Prostate cancer stage and grade at diagnosis are important predictors of survival and may also have different etiologic factors. To account for the possibility that such clinical subtypes may have, at least in part, different genetic susceptibility genes, we considered two case phenotypes and made a joint analysis of these two phenotypes and the control

group. This procedure should maintain statistical power to detect susceptibility loci that are specific to one subtype while preserving most of the power to detect SNPs that are common to both subtypes.
Thus, case phenotypes were categorized in two levels:

1. ***Non-aggressive*** prostate cancer patients: patients with non-invasive cancer (stage $\leq$2) and low Gleason score (score $\leq$6) tumor.
2. ***Aggressive*** prostate cancer patients: patients with invasive cancer (stage $\geq$3) or high Gleason score (score $\geq$7) tumor.

Analytic approaches
We present results from two distinct analytic approaches. The first scheme is more frequently used in case control studies. The second scheme takes full advantage of the prospective nature of the PLCO cohort and the incidence density sampling.

Analysis with Single selection
For this scheme, which will be more familiar to non-epidemiologists, does not account for the dynamic nature of the cohort. Genotypes of individuals that have been selected as a case in the relevant phenotype case group are counted once as a case and never as a control. Individuals who have been selected several times as controls but had not developed prostate cancer during follow-up are counted only once in the control group.

Analysis accounting for incidence density sampling
Selection of controls from cases identified in a cohort that accounts for the dynamic nature of the cohort including development of disease during follow-up and timing of entry to and exit from follow-up may have more power to detect an association than the Single selection method. The main feature of incidence density sampling, as used for control selection here, is that controls are selected independently for each case among those who are at risk at the time of the diagnosis of the case; i.e., among those who would become a case in the study had they developed disease at the same time. Independent selection means independent of future case status, of entry and exit times, and of selection as a control for other cases. Thus individuals may be selected several times as controls or as case subsequent to being included as a control. Genotypes of individuals that have been selected multiples times are taken into account the same number of times as selected, each time in the appropriate phenotypic group (Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. III. Design options. *Am J Epidemiol* 1992;135:1042-50).

Genotypes
In order to maintain high power to detect SNPs that are involved in non multiplicative models (such as complete recessivity or over-dominance), we provide analyses of the data based on genotype frequencies. Each of the three possible genotype states are considered separately. Accordingly, for autosomal loci, analysis of each case phenotype uses a statistical test with two degrees of freedom for each case phenotype considered separately (aggressive and non aggressive separately yield 4 degrees of freedom with 3 genotypes). For tests involving X-linked loci, a single degree of freedom is used per case phenotype.

Single SNP statistics
In order to expedite public access to the data, the first-pass analysis of the CGEMS data aims at detecting association of single SNPs with prostate cancer susceptibility. Multi-SNP approaches, such as haplotype association, have not yet been performed.

Statistical tests.

We performed four sets of analyses. For each test, analysis included

- 301,671 **SNPs**,
- 491 ***cases*** diagnosed with ***non-aggressive*** tumors and
- 691 ***cases*** diagnosed with ***aggressive*** tumors.

.
The characteristics of the four tests are:

1. ***Single selection, unadjusted*** score test
   - 1,105 ***controls*** that were not diagnosed with prostate cancer at the start of follow-up.
   - ***3-by-3 contingency table*** of genotypes by phenotypes was constructed.
   - ***No adjustment*** for covariates.
   - The ***p-value*** from the standard test of independence was computed from this table, based on a chi-squared test with up to 4 degrees of freedom.

2. ***Single selection, adjusted*** score test
   - 1,105 ***controls*** that were not diagnosed with prostate cancer at the start of follow-up.
   - ***Polytomous logistic regression*** was performed.
     - The outcome variable are the 3 phenotypes
     - Regression variable was indicator variable for genotype
   - ***Adjustment*** for
     - age group at randomization (4 groups),
     - region of recruitment (9 regions),
     - and an indicator variable for cases diagnosed within one year of entry to the trial.
   - The ***p-value*** was obtained from a score test with up to 4 degrees of freedom.

3. ***Incidence-density sampling, unadjusted*** score test
   - 1,174 ***controls*** using an incidence density sampling strategy.
   - ***3-by-3 contingency*** table of genotypes by phenotypes was constructed.
   - ***No adjustment*** for covariates
   - The ***p-value*** from the standard test of independence was computed from this table, based on a chi-squared test with 4 degrees of freedom (or fewer if there were empty cells).

4. ***Incidence density sampling, adjusted*** score test
   - 1,174 ***controls*** selected using an incidence density sampling strategy.
   - ***Polytomous logistic regression***
     - The outcome variable are the 3 phenotypes
     - Regression variable was indicator variable for genotype
   - ***Adjustment*** for
     - age group at randomization (4 groups),
     - region of recruitment (9 regions),
     - and an indicator variable for cases diagnosed within one year of entry to the trial.
   - The ***p-value*** was obtained from a score test with up to 4 degrees of freedom.

Interpreting the results

In examining the results one should keep in mind the following points:

1. Markers were selected on genomic criteria, not on functional basis. In the absence of complementary information, each of the SNPs has a low *a priori* probability. Observation of a low p-value in these tables is not sufficient evidence to demonstrate an association for the marker; additional studies are required to confirm the association. For this analysis, we expected to observe roughly $\alpha \times 3 \times 10^5$ p-values lower than a specified α when there is one statistical test for each of 3×10^5 SNPs by chance alone; thus for $\alpha = 10^{-3}$ or $\alpha = 10^{-5}$, we expected to observe 300 and 3 SNPs, respectively, meeting the criterion by chance. In the pre-computed analysis presented we observed 314±13 (with a range of 301 to 327) depending on which of the four tests was selected for $\alpha = 10^{-3}$. For $\alpha = 10^{-5}$, we observed between 7 and 9 SNPs for each of the four tests. Nevertheless, the observation of a low p-value for a SNP in this GWAS alone does suggest that the associated gene or chromosomal region has an increased likelihood of harboring a prostate cancer susceptibility locus but follow-up analysis is required and is planned in the follow-up phases of CGEMS (http://cgems.cancer.gov/).

2. Many pairs of SNP markers may have substantial correlation between them. In fact, correlation may extend across several markers on the same chromosomal region. Before interpreting the observation of clustering of SNPs with low p-values in a small chromosomal region as a strong signal of the presence of susceptibility loci in the region, one must consider that the clustering may be a consequence of linkage disequilibrium between neighboring SNPs. Similarly, the p values across the 4 statistical tests are highly correlated.

3. The four tests we used for each SNP are strongly associated. It is probably best to choose one test for exploratory purposes. We recommend using the fourth one, ***Incidence density sampling, adjusted*** score test for exploratory purposes.

Citation of data used:

Please cite the website for publications related to data available on this website (http://cgems.cancer.gov/) and reference the full name of the study, Cancer Genetic Markers of Susceptibility.

Appendix for statistical test

COMPUTATION OF 4 D.F SCORE-TEST FOR CGEMS ANALYSIS

The association of each SNP with advanced and non-advanced prostate cancer was tested using a 4 d.f score-test based on a polytomous logistic regression model[1]. If $Y = 0, Y = 1$ and $Y = 2$ denote controls, non-advanced and advanced prostate cancer cases, respectively, we specify the probability of observing a phenotype category as a function of the genotype data (G) and a set of co-factors (Z) as

$$\begin{aligned} &\Pr(Y = 0 \mid G, Z) = K \\ &\Pr(Y = 1 \mid G, Z) = K \exp\{\alpha_1 + \gamma_1 Z + \beta_{11} I(G = 1) + \beta_{12} I(G = 2)\} \\ &\Pr(Y = 2 \mid G, Z) = K \exp\{\alpha_2 + \gamma_2 Z + \beta_{21} I(G = 1) + \beta_{22} I(G = 2)\} \end{aligned} \qquad (0.1)$$

where K denotes a normalizing constant, $I(G = 1)$ and $I(G = 2)$ denote indicator variables for heterozygous and homozygous variant genotypes for a given SNP. In (1.1), the parameters β_{11} and β_{12} denote the log-odds-ratios associated with heterozygous and homozygous variant genotypes for non-advanced cases and β_{21} and β_{22} denote those for advanced cases. The null hypothesis of interest is

$$H_0 : \beta_{11} = \beta_{12} = \beta_{21} = \beta_{22} = 0$$

The parameters α_1 and α_2 determine the baseline probability of non-advanced and advanced prostate cancer for subjects with homozygous wild-type genotype $(G = 0)$ and some reference values for the covariates. The parameters γ_1 and γ_2 denote the log-odds-ratio associated with covariates (Z) for advanced and non-advanced cases, respectively. We implemented an Iterated Re-weighted Least Square (IRLS) algorithm[2] for obtaining the maximum-likelihood estimates of the parameters from model (1.1).

Steps for score-test calculations

1) Obtain estimates of covariate effects under the null

Estimate of $\theta_1 = (\alpha_1, \gamma_1)$ and $\theta_2 = (\alpha_2, \gamma_2)$ using the IRLS algorithm by setting the design matrix as $X = [1\ Z]$

2) Compute the score-vector

The formula for score-function S_{ij} for β_{ij} is given by

$$S_{ij} = \sum_{m=1}^{N_0+N_1} I(G_m = j)\{I(Y_m = i) - p_i(Z_m)\},$$

where $p_i(Z) = \Pr_{\theta_i}(Y = i \mid Z)$ is computed using estimate of θ_i obtained from step (1).

Alternatively, the score can be written in terms of "cell frequencies" (as opposed to individual level data) as following. Suppose the covariate Z defines a total of K strata. Let n_{ijk} and N_{jk} be the number of subjects in the data with $(Y = i, G = j, Z = z_k)$ and $(G = j, Z = z_k)$, respectively. Then the score S_{ij} can be written as

$$S_{ij} = \sum_{k=1}^{K} \{n_{ijk} - n_{+jk}\, p_i(z_k)\},$$

which has the usual (O-E) form. Define the score vector $\boldsymbol{S} = (S_{11}, S_{12}, S_{21}, S_{22})$.
Now we need to find the variance of the score-vector.

3) Variance calculation

Steps for variance calculations

3.1) Define the design matrix $X = [1\ Z\ I(G=1)\ I(G=2)]$

3.2) Based on this design matrix and parameter values $(\theta_1 = \hat{\theta}_1, \beta_{11} = 0, \beta_{21} = 0, \theta_2 = \hat{\theta}_2, \beta_{21} = 0, \beta_{22} = 0)$ compute the information matrix which can be as part of the IRLS algorithm.

3.3) Get $V = I^{-1}$ and extract the 4 by 4 sub-matrix from V, say $V_{\beta\beta}$ that corresponds to the rows and columns for the four β_{ij} parameters. The final score-statistics can be computed as

4) The final test-statistics is now given by $T = \boldsymbol{S} V_{\beta\beta} \boldsymbol{S}'$

Handling missing data on genotypes

The null model in step (1) should be fitted using all subjects that have covariate data Z, irrespective of whether those subjects have missing genotypes or not. Note that, under the null, this gives statistically the most efficient estimate of the covariate effect parameters θ_1 and θ_2. Moreover, this will be computationally also very efficient as one has to fit the "null" polytomous regression model only once. Alternatively, for computing the test of association for a particular SNP, one can estimate the parameters of the null model using only the data from those subjects who have complete genotype data for that SNP, but this will require fitting up to 500K different "null" polytomous model.

Once θ_1 and θ_2 are estimated using the full data, the score-vector calculation for a particular SNP will remain as above, except that the "sums" would now involve only those subjects who has complete genotype data on that particular SNP. The information matrix (I) calculation for a particular SNP will similarly involve only those subjects with complete genotype data for that particular SNP, except that the sub-matrix of I that corresponds to the parameters (θ_1 and θ_2) should be computed based on all the subjects that went into estimation of these parameters. This sub-matrix is simply given by the information matrix computed in step (1).

References

(1) Hosmer, D. and Lemeshow, S (1989). Applied logistic regression, John Wiley and Sons, New York.
(2) Generalized linear model, 2nd Edition. 1989. McCullagh P. and Nelder J.A. Chapman and Hall, London.